Inappropriate Content Detection and Categorization Approach

Nabil Badri, Ferihane Kboubia, Anja Habacha Chaibi

RIADI Laboratory, ENSI School, University of Manouba, 2010 La Manouba, Tunisia

Deep Learning Indaba 2023, Accra, Ghana, 3rd to 9th September 2023

Abstract

Over the past decade, increased use of social media has led to an increase in hate content. To address this issue new solutions must be implemented to filter out this kind of inappropriate content. Because manual filtering is difficult, several studies have been conducted in order to automate the process. The objective of this thesis is to provide an overview of the detection and classification of abusive and hate speech, including the challenges and approaches to address them.

Introduction

With the widespread of internet, huge increase in smartphone usage rate in recent years, the freedom of expression privilege granted after the Tunisian revolution, the mask of anonymity that the internet provides, and despite the fact that Twitter's terms of use forbid such inappropriate content, the spread of derogatory and hate speech has increased. It became easy to spread inappropriate content on social media, such as Twitter and Facebook, against individuals or groups. Furthermore, toxic language can take various forms, such as cyberbullying, which was one of the major reasons behind suicide.

Abusive and hate speech is a pervasive problem in society, with far-reaching implications. It is essential to be able to detect and classify these types of speech in order to effectively combat it.

This presentation will provide an overview of the detection and classification of abusive and hate speech, including the challenges and approaches to address them.

Definition of Hate Speech

"It covers all forms of expression which spread, incite, promote or justify racial hatred, xenophobia, anti-Semitism or other forms of hatred based on intolerance, including: intolerance expressed by aggressive nationalism and ethnocentrism, discrimination and hostility against minorities, migrants and people of immigrant origin. "

[Committee of Ministers of the Council of Europe]

Challenges

✓Detecting and classifying abusive and hate speech is a difficult task due to the complexity of language. It is also difficult to identify the intent behind the speech, which can make it difficult to determine if it is truly abusive or hateful.

✓ In addition, different types of speech can be interpreted differently by different people, making it difficult to accurately classify the speech.

✓The use of automated techniques for detecting and classifying abusive and hate speech has the potential to be both beneficial and problematic.

✓ On the one hand, it can help identify and mitigate such behavior, but on the other hand, it can also lead to false positives and other errors.

✓In addition, there are also challenges associated with training and deploying these techniques.

✓For example, it can be difficult to create datasets that accurately reflect the types of speech that need to be detected and classified.

✓ Furthermore, deploying these techniques at scale can be computationally expensive and difficult to manage.

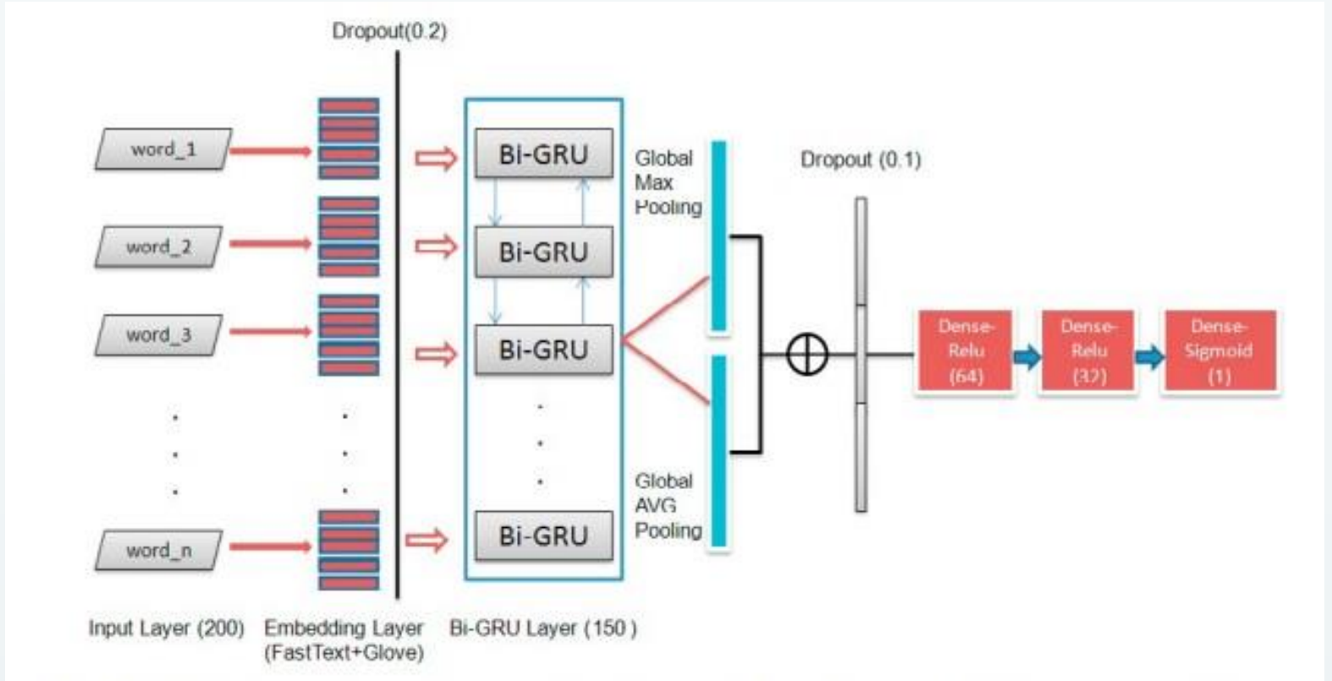
✓Yes, detection and classification of abusive and hate speech is a difficult task but by using natural language processing and machine learning techniques, it is possible to detect and classify these types of speech.

Approaches

Two approaches to detecting and classifying abusive and hate speech is to use:

✓NLP (Natural Language Processing) techniques can be used to analyze the language and identify patterns that indicate abusive or hateful speech.

✓In addition, machine learning algorithms can be used to classify the speech, by training the algorithm on a dataset of examples of abusive and hate speech.



The Bi-GRU model with combination of FastText and Glove embedding.

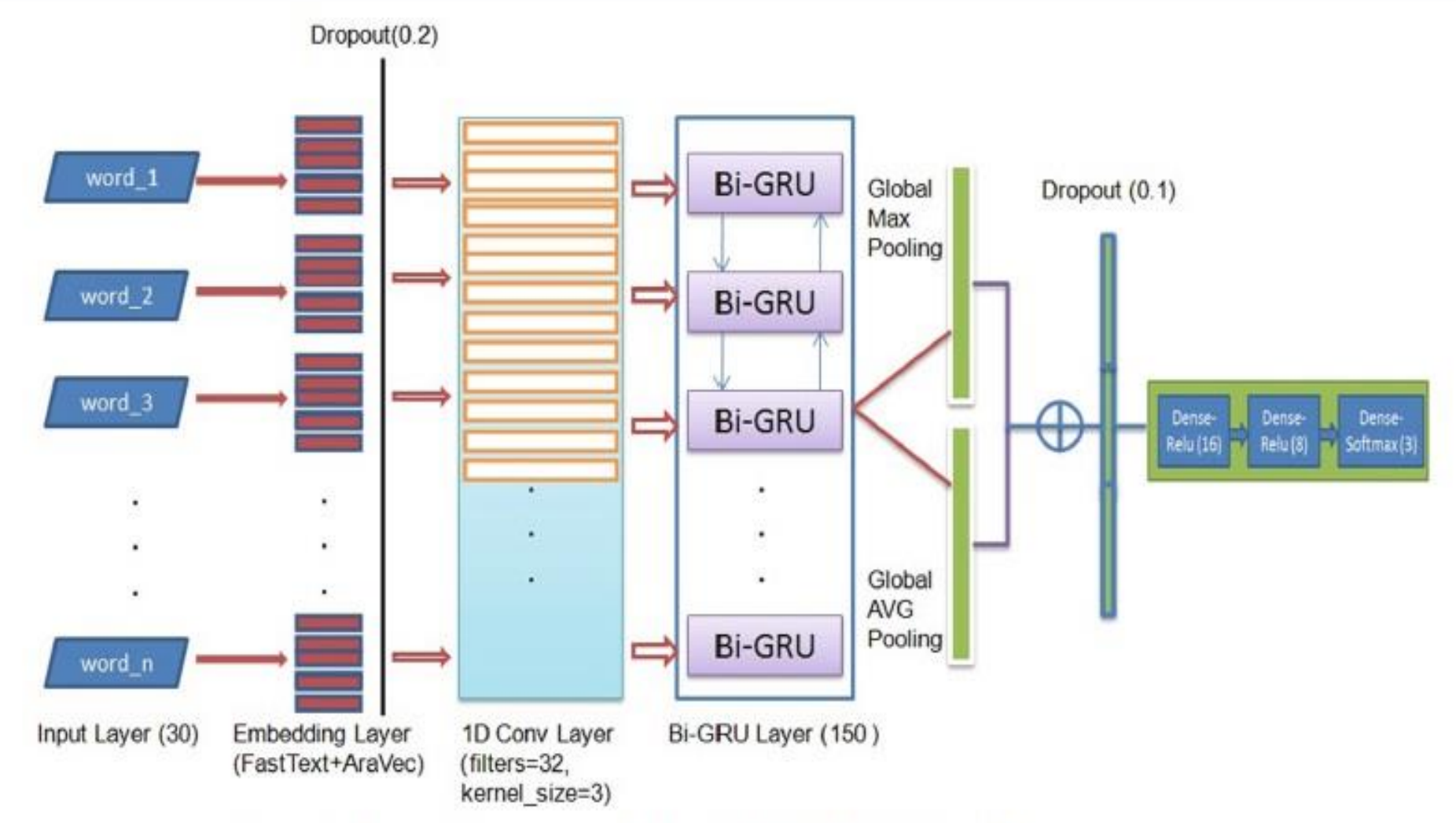
Natural Language Processing

•Natural language processing (NLP) is a technique used to analyze text and extract meaning from it.

• It can be used to detect abusive and hate speech by analyzing the words and phrases used in a given text.

•NLP can also be used to classify the type of speech, such as whether it is offensive, aggressive, or simply negative.

•NLP can be used to detect and classify abusive and hate speech with a high degree of accuracy. However, it is limited in its ability to identify context and nuance, which can be important in determining the intent behind a given statement.



Adopted architecture of the CNN-BiGRU model.

Machine Learning

- Machine learning is a type of artificial intelligence that can be used to detect and classify abusive and hate speech.
- Machine learning algorithms can be trained to recognize patterns in text and to classify them as abusive or non-abusive.
- These algorithms can also be used to identify the type of speech, such as whether it is offensive, aggressive, or simply negative.
- Machine learning has the potential to be more accurate than NLP in detecting and classifying abusive and hate speech.
- However, it is also more computationally expensive, and it can be difficult to train the algorithms to recognize subtle nuances in language.

Used datasets

Table 1. OLID Dataset.

Classes	Train	Test	Total tweets
Offensive	4400	240	4640
Not Offensive	8840	620	9460
	13240	860	14100

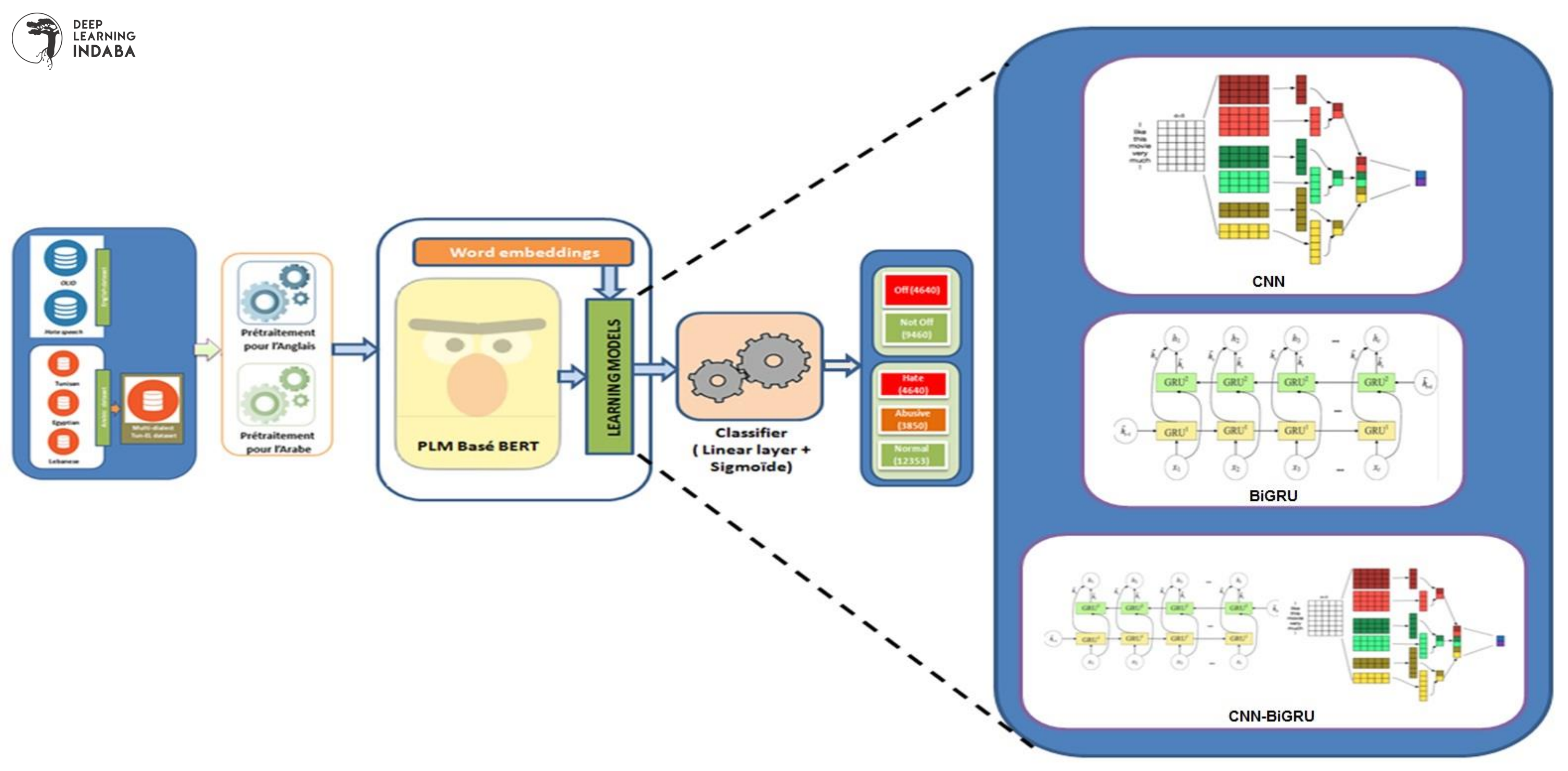
Table 2. Hate speech Dataset.

Classes	Train	Test	Total tweets
Normal	11213	2817	53851
Abusive	21781	5369	27150
Spam	43003	10848	14030
Hateful	3999	966	4965
	79996	20000	99996

List of datasets used for the study of hate speech in the Arabic language.

Ref.	Source	Classes	# Comments	M tongue
D1 [11]	F & YT	Hate, Normal, Abusive	6,040	Tunisian
D2 (*)	F	Hate, Normal, Abusive	437	Tunisian
D3 [14]	T	Hate, Normal, Abusive	5,846	Lebanese
D4 [16]	T	Abusive, hateful, Offensive, Disrespectful, Fearful, Normal	3,354	Lebanese
D5 [13]	T	Obscene, Offensive, Clean	1,101	Egyptian

Legend: 'M tongue' = Mother tongue, 'F' = Facebook, 'YT' = Youtube, 'T' = Twitter.
(*): We collected and annotated this dataset.



Hate speech classification system expanded on the left side of the figure and zoomed in learning models on the right side

Other Techniques

In addition to natural language processing and machine learning, there are other techniques that can be used to detect and classify abusive and hate speech.

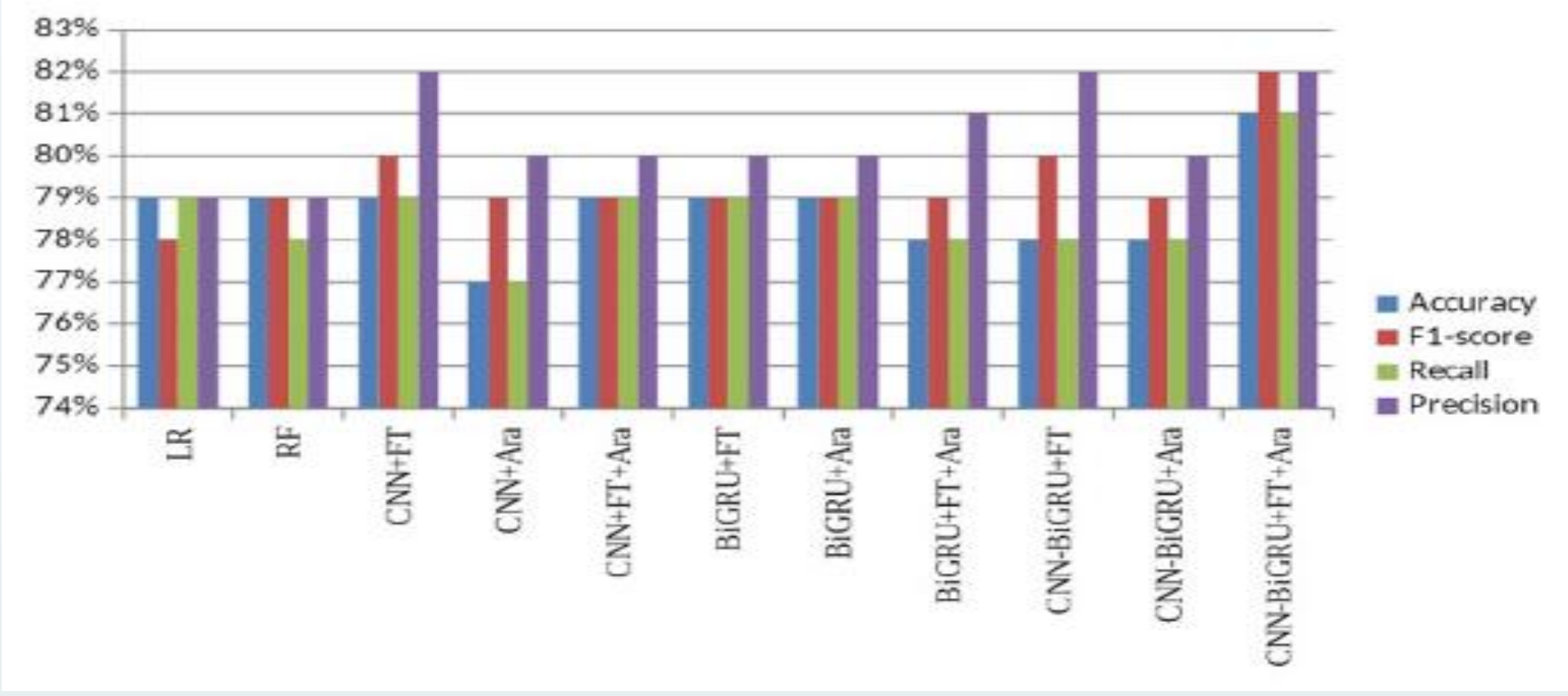
- These include lexical analysis, which looks at the words and phrases used in a given text, and sentiment analysis, which looks at the tone and sentiment of the text.
- These techniques can be used in combination with NLP and machine learning to increase the accuracy of detecting and classifying abusive and hate speech.
- However, they are also limited in their ability to identify context and nuance, which can be important in determining the intent behind a given statement.

Results and Applications

OLID Dataset (2 classes)

Models	Acc	Not P	Not R	Not F1	Off P	Off R	Off F1	Weighted Average P	Weighted Average R	Weighted Average F1	F1-M
Baseline RoBERTa	**0.83**	0.84	**0.94**	**0.89**	**0.78**	0.53	0.63	0.82	**0.83**	0.81	0.76
Baseline BiGRU & Fasttext (FT) Embedding Only	**0.83**	**0.86**	0.92	**0.89**	0.74	**0.61**	0.67	**0.83**	**0.83**	**0.83**	**0.78**
Baseline BiGRU & Glove Embedding Only	**0.84**	**0.86**	0.93	**0.89**	0.76	**0.61**	**0.68**	**0.83**	**0.84**	**0.83**	**0.78**
BiLSTM (Marcos Zampieri,2019)	-	0.83	**0.95**	**0.89**	**0.81**	0.48	0.60	0.82	0.82	0.81	0.75
SVM (Marcos Zampieri,2019)	-	0.80	0.92	0.86	0.66	0.43	0.52	0.76	0.78	0.76	0.69
Our experiments: BiGRU model with GloVe and Fasttext embedding (BiGRU_Glove_FT)											
BiGRU_Glove_FT	**0.84**	**0.87**	**0.93**	**0.90**	**0.77**	**0.63**	**0.69**	**0.84**	**0.84**	**0.84**	**0.79**

Final results of the baselines and our experiments, for the OLID datasets. The **bold** figures represent the best scores. Legend: 'F1-M' = F1-Macro.



Performance of the CNN, BiGRU and CNN-BiGRU models on Tun-EL dataset

-- The detection and classification of abusive and hate speech can be used in a variety of applications.

✓To moderate online conversations

✓It can be used, or to detect and flag abusive or hateful content on social media.

✓It can be used to identify and remove abusive or hateful content from websites

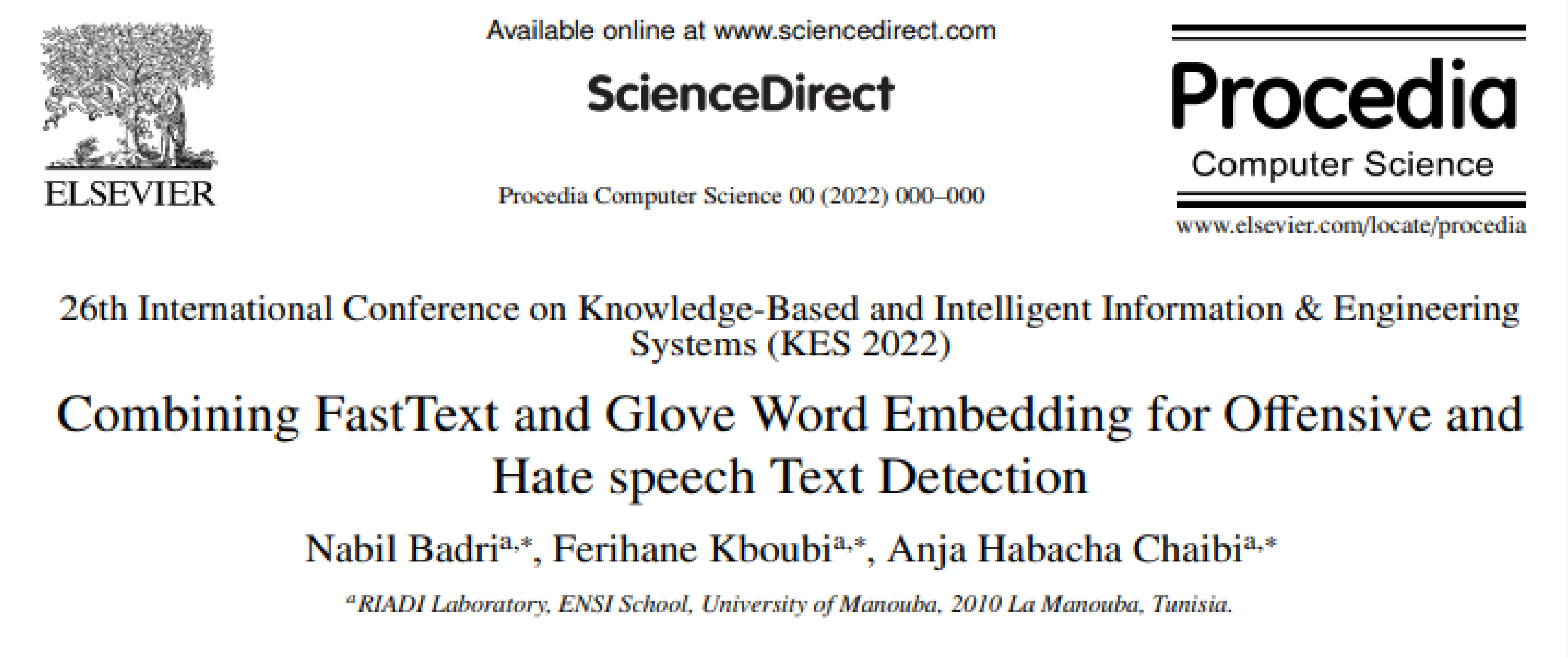
✓To monitor and analyze online conversations for signs of abuse or hate.

Conclusion

✓The detection and classification of abusive and hate speech is a difficult but important task.

✓Using natural language processing and machine learning techniques, it is possible to detect and classify these types of speech.

✓These techniques can then be used in a variety of applications, such as moderating online conversations and identifying and removing abusive or hateful content from websites.



Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Procedia Computer Science 00 (2022) 000–000

Procedia Computer Science

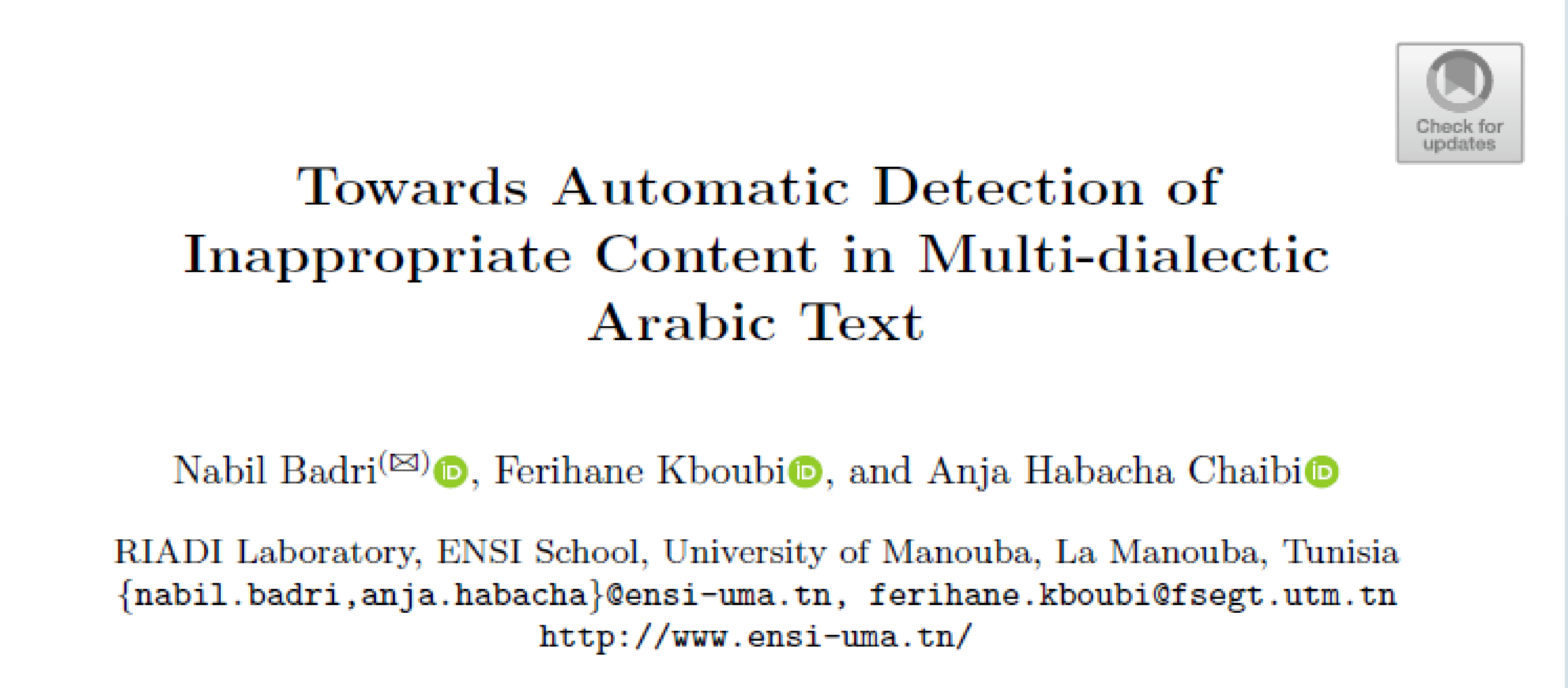
www.elsevier.com/locate/procedia

26th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2022)

Combining FastText and Glove Word Embedding for Offensive and Hate speech Text Detection

Nabil Badri[a,*], Ferihane Kboubi[a,*], Anja Habacha Chaibi[a,*]

[a]*RIADI Laboratory, ENSI School, University of Manouba, 2010 La Manouba, Tunisia.*



Check for updates

Towards Automatic Detection of Inappropriate Content in Multi-dialectic Arabic Text

Nabil Badri(✉), Ferihane Kboubi, and Anja Habacha Chaibi

RIADI Laboratory, ENSI School, University of Manouba, La Manouba, Tunisia
{nabil.badri,anja.habacha}@ensi-uma.tn, ferihane.kboubi@fsegt.utm.tn
http://www.ensi-uma.tn/